Supplement to “Revisiting the Impacts of Teachers”

Jesse Rothstein*

January 2017

This note contains supplementary material that could not be included in the published version or online appendix of “Revisiting the Impacts of Teachers” (Rothstein, 2017).

Section 1 responds to the arguments made in Chetty, Friedman, and Rockoff’s (hereafter, “CFR”) Reply, CFR (2016).

I respectfully disagree with many of the conclusions drawn by CFR (2015), which in many cases are based on claims that are theoretically correct but turn out, upon investigation, to be empirically irrelevant. None of the evidence presented by CFR (2015) alters the main conclusions of my earlier draft, which persist in the current version:

1. That the CFR (2014a; hereafter, “CFR-I”) research design is not a valid quasi-experiment because the treatment is correlated with observable determinants of the outcome;

2. That much but not all of the problem derives from CFR-I’s exclusion of a non-random subset of classrooms from school-grade-subject-year means;

3. That estimates that adjust for differences in observables indicate a non-trivial but not enormous degree of “forecast bias”; and

4. That estimates of teachers’ long-run effects are not at all robust and quite likely to be biased by student sorting.

Section 2 presents some important specification and robustness analyses, focusing on the treatment of teachers who are observed for only one or two years and

*Goldman School of Public Policy and Department of Economics, University of California, Berkeley. E-mail: rothstein@berkeley.edu.

who therefore lack leave-two-out value-added (VA) predictions. These demonstrate clearly that my results do not derive from mispecification of the model used to predict these teachers' VA, and are robust to a variety of choices about how the are handled so long as *something* is done to avoid the sample selection bias in CFR-I's main specification.

1 Rejoinder to CFR (2015)

The exchange between myself and Chetty, Friedman, and Rockoff (CFR) has involved several rounds of private communication, dating back to 2010, and a more recent exchange of public drafts and responses. Throughout, it has been constructive and scholarly, and I have learned a great deal from it. I am grateful to CFR for their role in it, and the final version of my Comment reflects many good points that CFR have made.

Nevertheless, CFR and I continue to have sharply different interpretations of what the empirical patterns mean for the substantive questions under investigation. My Comment reflects my interpretation; CFR offer a very different interpretation in their Reply. In this appendix, I discuss the December 2016 version of CFR's Reply (CFR 2016), written in response to a version of my Comment (Rothstein, 2016) that differs only cosmetically from the final, January 2017 version. To ensure a complete record, the original, March 2016 version of my rejoinder (which responded to the July 2015 version of CFR's Reply) will remain posted on my webpage, at `http://eml.berkeley.edu/~jrothst/CFR/supplement_mar2016.pdf`.

I respectfully disagree with many of the conclusions drawn by CFR (2016), which in many cases are based on claims that are theoretically correct but turn out, upon investigation, to be empirically irrelevant. None of the evidence presented by CFR (2016) alters the main conclusions of my earlier draft, which persist in the current version:

1. That the CFR-I (2014a) research design is not a valid quasi-experiment because the treatment is correlated with observable determinants of the outcome;

2. That much but not all of the problem derives from CFR-I's exclusion of a non-random subset of classrooms from school-grade-subject-year means;

3. That estimates that adjust for differences in observables indicate a non-trivial but not enormous degree of "forecast bias"; and

4. That estimates of teachers' long-run effects are not at all robust and quite likely to be biased by student sorting.

I begin by laying out CFR (2016)'s six main arguments, in order of their importance to my conclusions, along with my responses. I follow this by presenting simulation evidence to support one of these responses. In the interests of space, I do not discuss other arguments made in CFR's response that are less relevant to my conclusions.

CFR (2016)'s six main arguments are:

1. *Examination of prior test scores is not informative about the validity of CFR-I's quasi-experimental research design, because value-added is estimated from prior test scores and is thus mechanically correlated with them.*

It is theoretically correct that the use of prior test scores in the construction of the VA measures could create a spurious correlation, making it appear that changes in teacher VA are not randomly assigned. But in practice, this does not account for the result. The main text and Appendix B present a number of analyses that probe this possibility. All indicate that the failure of the placebo test is real, not spurious. The most definitive is an alternative placebo test that is based solely on non-test student characteristics (race, gender, special education, free lunch status, limited English status, grade repetition, etc.). This test is entirely immune from mechanical correlations, but also shows that changes in mean teacher VA, as estimated by CFR-I, are significantly related to changes in student preparedness (see Table 2).

2. *The primary source of the correlation between changes in teacher value added (VA) and changes in prior test scores is common shocks that affect both. When these so-called "mechanical effects" are addressed via changes in the specification, the correlation is eliminated.*

CFR (2014c; 2014d; 2015; 2016) have advanced this idea in a series of public responses over the last several years, pointing to potential mechanical effects deriving from teachers who follow students across grades or from school-year-subject-level shocks. As noted above, explanations based on test score dynamics cannot possibly account for the placebo test result, as it holds even when non-test variables are used in place of prior test scores. Moreover, for each proposed mechanical channel, I have implemented alternative specifications of the placebo test that close off that channel. In particular, I close off the teacher-follower channel by instrumenting with VA changes computed only over non-follower teachers, and I close off the school-year-subject shock channel by using "leave three out" VA measures that do not rely on data from $t-2$ in computing VA predictions for $t-1$ or t. Results are remarkably stable across specifications (see Appendix Table B1).

CFR (2016) suggest that there may be school-level shocks that are correlated across years, so that shocks in $t-3$ influence both VA predictions for $t-1$ teachers (even when $t-2$ data are excluded) and the prior year scores of $t-1$ students, which are measured in $t-2$. Serially correlated school-level shocks could produce the failure of my placebo test even when I use leave-three-out VA scores that do not rely on $t-2$ data.

To ensure that my results are not driven by this channel, I estimated specifications (Appendix Table B1) that exclude all data from several years before the $\{t-1, t\}$ window from the VA predictions. If in fact the placebo test result derived from serially correlated shocks, the coefficient should decline as more years are excluded. But in fact this has essentially no effect on the results – even when I base VA predictions solely on *future* data. Thus, while CFR-I present simulation evidence that serially correlated shocks *could* drive the results, the empirical evidence from real data indicates that they do not.

It is also worth noting that the dynamics that CFR (2016) propose as sources of mechanical effects would in general invalidate not just the placebo test but also CFR-I's quasi-experimental research design itself, and would lead CFR-I to understate forecast bias. School-year or school-subject-year shocks that are correlated between $t-2$ and $t-1$ would invalidate the design, as the leave-two-out teacher VA predictions for $t-1$ would be influenced by shocks correlated with those to students' $t-1$ test scores.[1] It would take a very particular dynamic structure to generate correlations between $t-3$ and $t-2$ scores but not between those in $t-2$ and $t-1$. Similarly, the presence of meaningful numbers of "follower" teachers would imply that the outcome in the quasi-experiment reflects not only the quality of the grade-g teachers but also the (correlated)

[1] CFR (2016) present a specification with school-subject-year FEs. But with only two or three observations (grades) per school-subject-year cell, these specifications rely very heavily on a strict exogeneity assumption that is prima facie violated by teachers who switch grades within schools. In my explorations with simulated data – including with the data generating process of the simulations used in CFR (2016)'s Table 3 – I have found that these specifications are very poorly behaved.

quality of grade $g-1$ teachers, and thus that the quasi-experimental coefficient overstates the parameter of interest, λ.

3. *The augmented quasi-experimental specification that includes a control for the change in prior year scores yields a biased estimate of the forecast bias coefficient* λ.

Again, this is theoretically possible, but the claim that it is relevant in practice is pure speculation unsupported by evidence. CFR (2016) hypothesize that the change in prior year scores has two components, with one component correlated with the change in VA but not with the change in end-of-year scores and the other correlated with end-of-year scores but not with VA. This might be a reasonable hypothesis if the "mechanical effects" claims discussed above held up. Even here, quite restrictive dynamic structures would be needed to generate mechanical effects from sources that are uncorrelated with the dependent variable in CFR-I's analyses. CFR (2016) argue for "non-parametric" specifications, but their specifications and simulations generally rely on quite strong implicit assumptions. As noted above, the evidence does not support CFR's claims about mechanical effects. Without them, while anything is possible, the only reasonable conclusion is that CFR's (2016) conclusions rely on quite speculative, unsupported assumptions.

It is also possible, and more likely, that both the specification without a control for prior year scores (as in CFR-I) and one with such a control (as in my preferred analyses) are biased by unmeasured components of the endogeneity of teacher VA changes. I do not claim that the specification with controls is highly credible. But in the presence of clear evidence that the quasi-experimental treatment is not randomly assigned, and that this is *not* attributable to CFR (2016)'s hypothesized mechanical effects, a specification with controls is preferable, in my view, to one that does nothing to address the endogeneity of treatment. Moreover, I show (see Table 3) that the top-line result of forecast bias around 10-15% (i.e., of $\hat{\lambda}$ around 0.85-0.9) is robust to several ways of addressing the endogeneity, which adds to my confidence in the result.

4. *An analysis restricted to school-grade-subject-year cells without missing data is the most definitive way to address concerns about sample selection due to missing data, and validates CFR-I's conclusion that VA scores are forecast unbiased.*

I disagree that this is the most definitive way to address concerns about sample selection due to missing data – it requires discarding between three-quarters (New York) and four-fifths (North Carolina) of the school-grade-subject-year cells, and estimates are quite imprecise. Moreover, the remaining sample includes fewer teachers who are new to teaching or to the sample grades, and forecast bias in this subsample might be different from that in the broader population.

More importantly, as discussed in Section 2, below, the subsample analysis does not validate the conclusion of no forecast bias. First, I find that the placebo test coefficient is quite large and statistically significant even in the complete data subsample. Second, CFR-I inexplicably drop the school-year fixed effects from their preferred specification when they analyze the complete data subsample. When I include them the estimate of λ is 0.918 without controlling for prior year scores and 0.899 (and significantly different from one) when this control is included. This is broadly similar to what is obtained from the full sample.

Thus, at most this subsample analysis shows that not *all* of the problem with CFR-I's specification is attributable to their exclusion of a non-random subset of classrooms from school-grade-subject-year means. It does not demonstrate (or even point in the direction) that the design is valid, or that forecast bias is zero, even locally for the small subset of schools without missing data. CFR (2016)'s statement that "[t]his approach consistently yields estimates of forecast bias close to zero in the New York, North Carolina, and Los Angeles datasets" is incorrect as it applies to North Carolina, and the single specification that CFR have reported from their dataset is not enough to demonstrate the point there either.[2]

5. *The inclusion of all classrooms in the analysis, using grand mean imputation, generates downward-biased estimates of the key parameter λ.*

We are in agreement that analyses that include all classrooms are not definitive, but rest on the appropriateness of the model used to predict teachers' VA. I focus on specifications that use the grand mean because this is the strategy proposed by CFR, who use it throughout their analyses for some (most of CFR-I's

[2] Despite repeated requests, CFR have never reported estimates of the subsample analysis with school-year fixed effects, so it is not clear whether the same result holds in New York. Bacher-Hicks et al. (2014) also report estimates of the subsample analysis, but like CFR include only year fixed effects.

specifications) or all (one failed robustness test in CFR-I, and the main specifications of CFR-II) of the classrooms with missing data.[3] It is also consistent with CFR's prediction model (seen as an example of Empirical Bayes methods) for classrooms that have data.

That said, the claim that my use of grand mean predictions accounts for my results is incorrect. CFR (2016) are correct that positively correlated VA across teachers within schools could lead to attenuation with grand mean predictions.[4] But again, this theoretical point is not empirically relevant. Results of both the placebo test and the forecast bias estimation are robust to a variety of alternative prediction strategies, including some that are robust to non-independence of teacher VA within schools (which is the source of bias under grand mean predictions). See the discussion in Section 2, below. And even when I follow CFR-I's preferred strategy of excluding classrooms without teacher VA predictions, the results are quite clear that λ is less than one in any specification that does anything to address the endogeneity of changes in teacher VA (Table 3).

Three other points are worth noting about the imputation issue:

- CFR (2016)'s attenuation argument may help to explain why some of the placebo test coefficients are smaller when all classrooms are included than when they are not (see Table 2); it suggests that the failure to reject the placebo test null hypothesis in some all-classroom specifications should not be taken as support for the exclusion restriction.

- CFR (2016) present simulation results to demonstrate bias from grand mean imputation when teacher VA is correlated within schools. This simulation assumes that there are no differences across classrooms in students' prior achievement. My argument for the importance of accounting for classrooms with missing teacher VA was predicated on the empirical result that students' prior scores are positively correlated with teacher VA, so excluding a classroom has effects of the same sign on mean teacher VA

[3] Throughout all of their quasi-experimental analyses, CFR-I and CFR-II impute VA scores of zero for teachers observed in $t-1$ and t but not in other years. At issue is whether to apply the same imputation to teachers observed only in a single year, as is done in CFR-I's Table 5, Column 2 and throughout CFR-II, or to exclude these teachers and their students from the analysis, as is done elsewhere in CFR-I. I see no basis for viewing the grand mean as the correct prediction for the first group of teachers but not for the second, and CFR have never offered an explanation for this, nor have they defended the implicit – and demonstrably false – missing-at-random assumption needed to support excluding either group of classrooms.

[4] An earlier draft of my comment (Rothstein, 2014) presented estimates that used all classrooms on one side of the regression and a subset on the other in order to build intuition for the full-sample results. CFR (2015) quite reasonably objected that these specifications were not very informative. They have therefore been removed.

and mean student preparedness that bias the $\hat{\lambda}$ coefficient upward. It is thus not surprising that CFR's simulation shows no bias from excluding classrooms with missing VA, as it fails to include the relevant features of the real data. Where the real data are concerned, CFR (2016) may object to the particular imputation model proposed by CFR-I, but they do not dispute that excluding classrooms with missing data, as in CFR-I's main analyses, biases $\hat{\lambda}$.

- Finally, the data generating process for CFR (2016)'s simulation violates the exclusion restrictions that CFR-I require to identify λ, even with random assignment and complete data, as these restrictions rule out non-zero intra-school correlations. If the intra-school correlation is non-zero, the change in the average of unbiased predictions of individual teachers' VA is not an unbiased prediction of the change in the average VA. If the correlation is positive, CFR-I's methods will likely overstate the change in VA, even with complete data, biasing $\hat{\lambda}$ upward.[5] This could offset bias from endogenous teacher switching (or from endogenous sample selection).

One final point: While we agree that specifications that include all classrooms rest on the appropriateness of the model used to predict teachers' VA, it is also true that specifications, like those that CFR-I prefer, which exclude a non-random set of classrooms also rest on assumptions. These assumptions are quite implausible – they require that student preparedness be uncorrelated with teacher VA. It is empirically the case that students' observables *are* correlated with teacher VA; whether their unobservables are as well is the entire point of the CFR-I exercise. So while it is reasonable to disbelieve specifications that rely on imputations, it is not reasonable to treat those that simply exclude teachers with missing data as unbiased.

6. *There exist assumptions in which CFR-II's two-step procedure for estimating the association between teachers' test score effects and their impacts on longer-run outcomes is consistent but the OLS regression with controls that I emphasize is not.*

[5] In CFR (2016)'s simulation, teachers' VA is known with certainty, so λ is identified even with non-zero correlations. But in CFR-I's actual data, a teacher's VA in one year is predicted based on noisy measures of her performance in other years. In this case, if true VA is correlated among teachers at the same school, the change in the school-grade-subject-year average predicted VA, ΔQ_{sgmt}, is a biased prediction of the change in the average true VA, violating CFR-I's Assumption 3.

This point responds to an earlier version of my comment (Rothstein, 2014). CFR (2015)'s discussion of this issue clarified it substantially for me, and the revised comment has been rewritten with this in mind.[6] I believe that the main point stands.

CFR are correct that the exclusion restrictions under which my approach identifies κ do not strictly nest those under which CFR-II's approach identifies that parameter. But they do not state clearly the conditions under which their approach is consistent but mine is not. This could occur only in very particular, implausible circumstances. In order for their approach to be consistent, one of two conditions would need to hold: Either observed student characteristics that are predictive of long-run outcomes must be uncorrelated with teachers' estimated value-added, or the between-teacher and within-teacher associations between these characteristics and students' long-run outcomes must be identical. Each of these conditions is demonstrably false. Moreover, even if both were to hold, CFR-II also require a strong, untestable assumption that *unobserved* determinants of students' long-run outcomes are uncorrelated with teacher VA.

My proposed strategy relies on a subtly different assumption, that unobserved determinants of students' long-run outcomes are uncorrelated with teacher test score VA *conditional on observables.* This is not strictly nested within CFR-II's assumptions. Specifically, my assumption might be violated if parents were able to discern teachers' long-run impacts τ and to sort into classrooms on this basis. This would create correlations between τ and students' unobserved characteristics, and would make the probability limit of $\hat{\kappa}_X$ differ from κ.

It is worth considering, however, the specific conditions under which the OLS estimator with controls is inconsistent but the CFR-II two-step estimator is consistent. This requires that parents do not sort into classrooms on the basis of teachers' test score value-added μ, as this would make the CFR-II estimator inconsistent, but that they do sort on the basis of the component of teachers' long-run effects that is unpredictable from μ, $\tau - E[\tau|\mu]$. In this case, student observables would be correlated with the error term in a bivariate regression of τ on μ, so the multiple regression coefficient would not converge to the bivariate coefficient of interest.

This condition strikes me as exceedingly unlikely. Thus, I place little poste-

[6] In personal communication regarding the long-run analysis, CFR emphasized measurement error in teacher VA. Responding to this, I (Rothstein, 2014) presented IV specifications designed to eliminate attenuation due to measurement error in an explanatory variable, with zero impact on the results. CFR now point to a different dynamic, so I no longer emphasize the IV results.

rior weight on the possibility that the multiple regression coefficient κ_X is biased but the CFR-II two-step estimator is unbiased. Because the assumptions needed for CFR-II's estimator are demonstrably false, and because violations of them are likely to bias upward the estimate of the association between teachers' test score VA and their long-run impacts, I also place little weight on the possibility that both are unbiased. The plausible alternatives, in my view, are that κ_X is unbiased but CFR-II's estimates are biased, or that both are biased. I thus have little confidence that CFR-II's cross-sectional estimates of the association between teachers' test score VA and their long-run impacts are informative about the association of interest. As discussed in my comment, I have similarly little confidence in CFR-II's quasi-experimental estimates, which as in CFR-I are quite sensitive to controls for failure of the quasi-experiment. In my view, the question of whether teachers' test score VA is correlated with their long-run impacts (if indeed teachers vary substantially in their causal effects on students' longer-run outcomes) remains unresolved.

2 Teachers with missing leave-two-out predictions

CFR-I's key VA measure used in each paper is a "leave-two-out" forecast of a teacher's outcomes in year t or $t-1$ based only on data from prior to $t-1$ or after t.[7] This forecast can be seen as an Empirical Bayes prediction of the teacher's impact in $t-1$ or t, and by construction is an unbiased prediction of the VA score in that year. When teachers are observed only in $t-1$ or t, however, there is no other data on which to base this forecast. In most of their analyses, CFR-I exclude such teachers, and their students, from their calculation of school-grade-year means. Rothstein (2017) argues that this sample selection biases the key coefficient $\hat{\lambda}$ toward the null hypothesis of $\lambda = 1$. Following one specification in CFR-I and most of the analysis in CFR (2014b; "CFR-II"), he includes these teachers and their classrooms, assigning them a VA prediction equal to the grand mean.

The grand mean is an unbiased prediction of every teacher's VA, and is the logical extension of the Empirical Bayes methodology for CFR-I's leave-two-out predictions. But the relevant prediction for CFR-I's quasi-experimental analysis is of the school-grade-year mean VA, not that of the individual teacher. If VA is correlated across teachers within schools, then the average of unbiased forecasts

[7] I do not review the notation of CFR-I and Rothstein (2017) in detail here; readers are referred to those papers for this.

for each teacher is a biased forecast of the average VA at the school. Failure to account for this would create upward bias in both CFR-I's quasi-experimental coefficient $\hat{\lambda}$ and Rothstein's (2017) placebo test coefficient. Importantly, this bias arises even if leave-two-out forecasts are available for every teacher. Avoiding it would require shrinking teachers' observed performance toward the school mean rather than toward the grand mean, and using school average performance rather than the overall average to predict VA for teachers with missing leave-two-out data.

Table 1 explores alternative strategies for assigning VA predictions to teachers with missing leave-two-out data. Following CFR (2015), I use CFR-I's leave-two-out predictions for teachers for whom they are available in every specification in this table, though the above discussion suggests that the should be changed as well.

Panel A presents CFR-I's main regression of the year-over-year change in school-grade-subject mean test scores on the corresponding change in mean teacher predicted VA. Panel B presents Rothstein's (2017) placebo test, replacing the dependent variable with the change in mean *prior year* scores. Panel C augments the Panel A specification with a control for the change in mean prior year scores.

The first two columns reproduce estimates from Rothstein (2017) for context: Column 1 leaves the teachers with missing leave-two-out predictions and their students out of the school-grade-year means, while column 2 includes them using the grand mean for the teachers' VA predictions. When the teachers are left out, $\hat{\lambda} = 1.03$ (standard error 0.02) when students' prior scores are not controlled, and the null hypothesis of $\lambda = 1$ is not rejected. But the placebo test fails, with a highly significant coefficient of 0.14, and when students' prior-year scores are controlled the key coefficient falls to 0.93 (0.02) and the null hypothesis is rejected. When teachers with missing leave-two-out predictions are included, even the baseline specification in Panel A rejects the null hypothesis ($\hat{\lambda} = 0.90$, SE 0.02). The placebo test result is weaker but still significant, and the specification that controls for observables yields $\hat{\lambda} = 0.86$ (SE 0.02).

Columns 3-5 present results from other imputations. Column 3 uses the (appropriately shrunken) mean residual of all teachers at the school in all years other than $t-1$ or t to forecast the VA of teachers in those years who are not seen outside that window. This method would be robust to correlations among teachers at the same school. Column 4 uses the mean residual of all teachers across all schools who are observed for two years or less. This captures the possibility that the teachers with missing leave-two-out predictions may systematically differ from others. Finally, Column 5 uses the mean for such teachers at the same school, as in other cases using only data from outside the $t-1$ to t window.

Results are qualitatively similar across all of the different imputation models. In each case, the baseline specification in Panel A yields an estimated $\hat{\lambda}$ between 0.90 and 0.93, all significantly different from one. The placebo test fails regardless of the imputation used, with the models that use only same-school data indicating much larger placebo test violations. And when prior scores are controlled, the key coefficient falls to between 0.85 and 0.89, again always significantly different from one. It is clear that non-independence of teacher VA within schools cannot account for Rothstein's (2017) results.

Table 2 takes a different approach to the issue of missing leave-two-out predictions. Column 2 of CFR-I's Table 5 suggests a substantial degree of forecast bias when teachers with missing VA predictions are assigned the grand mean VA, and as Table 1 indicates the same is true in the North Carolina sample. But CFR (2016) point instead to Columns 3 and 4 of CFR-I, Table 5, reproduced for the North Carolina sample in Rothstein (2017), Table A5. These limit the sample to school-grade-subject-year cells with few (Column 3) or no (Column 4) missing VA predictions, and in each sample they indicate less forecast bias. CFR (2016) interpret this as evidence that the imputation algorithm accounts for the result in Column 2, and argue that the Column 4 result in particular indicates that VA predictions are unbiased, at least in the subsample of school-grade-subject-year cells with no missing VA predictions.

But this result is not at all robust. In particular, it evaporates when school-year fixed effects are added. These fixed effects are included in CFR-I's main specifications but omitted without explanation from their Table 5.

The odd numbered columns of Table 2 report the four specifications from CFR-I's Table 5. Note that the placebo test coefficients are quite large in these columns, though the models with controls in columns 1, 5, and 7 yield λ estimates that are not distinguishable from 1 (in large part because the models without controls yield λ estimates well in excess of 1).

As noted, these specifications, following CFR-I, include only year fixed effects, rather than the school-year effects included in the models that CFR-I prefer in the rest of their analysis. This raises the possibility of bias from unmodeled school trends. The even numbered columns of Table 2 add back the school-year fixed effects.[8] This change reduces the placebo coefficients, which become insignificant in columns 6 and 8. But it also reduces the forecast bias coefficients. CFR-I's preferred model, which limits the sample to cells with no missing data, yields a forecast bias coefficient of $\hat{\lambda} = 0.92$ without controls and 0.90 (significantly different from one) with a control for the change in prior year scores. This is broadly similar to what is obtained from the full sample.

References

Bacher-Hicks, A., T. J. Kane, and D. O. Staiger (2014): "Validating Teacher Effect Estimates Using Changes in Teacher Assignments in Los Angeles," Working paper 20657, National Bureau of Economic Research.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014a): "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 104, 2593–2632.

——— (2014b): "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 104, 2633–2679.

——— (2014c): "Prior Test Scores Do Not Provide Valid Placebo Tests of Teacher Switching Research Designs," Unpublished manuscript. Downloaded October 13, 2014 from `http://obs.rc.fas.harvard.edu/chetty/va_prior_score.pdf`.

——— (2014d): "Response to Rothstein (2014) on "Revisiting the Impacts of Teachers"," Unpublished manuscript. Downloaded from `http://obs.rc.fas.harvard.edu/chetty/Rothstein_response.pdf` on October 13, 2014.

——— (2015): "Measuring the Impacts of Teachers: Response to Rothstein (2014)," July. Unpublished manuscript. Downloaded July 27, 2015 from `http://obs.rc.fas.harvard.edu/chetty/va_response.pdf`.

——— (2016): "Measuring the Impacts of Teachers: Response to Rothstein (2016)," December. Forthcoming, American Economic Review.

[8] One might worry that the no-missing-data subsample in Column 7 is not large enough to permit any degree of precision with school-year fixed effects. But standard errors increase by less than 20% when these are added, much less than the increase (of nearly 100%) when cells with missing VA predictions are discarded.

Rothstein, J. (2014): "Revisiting the Impacts of Teachers," Unpublished manuscript, October.

——— (2016): "Revisiting the Impacts of Teachers," March. Unpublished manuscript.

——— (2017): "Revisiting the Impacts of Teachers," January. Forthcoming, American Economic Review.

Supplement Table 1. Assessing sensitivity of results to the imputation model

	Excluding classrooms missing teacher VA predictions	Including all classrooms, assigning to teachers with missing VA predictions:			
		Grand mean	School mean	Missing mean	Missing mean at school
	(1)	(2)	(3)	(4)	(5)
	Panel A: Quasi-experimental models without controls				
Change in mean teacher predicted VA	1.030	0.904	0.915	0.933	0.911
	(0.021)	(0.022)	(0.022)	(0.022)	(0.021)
	Panel B: Models for change in prior-year scores				
Change in mean teacher predicted VA	0.144	0.092	0.134	0.084	0.128
	(0.021)	(0.022)	(0.023)	(0.023)	(0.022)
	Panel C: Models for change in end-of-year scores, with controls for change in prior-year scores				
Change in mean teacher predicted VA	0.933	0.860	0.850	0.892	0.847
	(0.015)	(0.017)	(0.017)	(0.017)	(0.017)
Change in mean student prior year score	0.675	0.536	0.535	0.536	0.535
	(0.004)	(0.009)	(0.009)	(0.009)	(0.009)

Notes: Specifications in column 1, panels A-C are identical to those in Table 1, Column 2; Table 2, Column 1; and Table 3, Column 2, respectively. Successive columns include all classrooms in the dependent and independent variables, varying the VA prediction assigned to teachers who are excluded in column 1. In column 2, these teachers are assigned the grand mean of zero. In Column 3, the prediction is based on the shrunken leave-two-out mean at the same school. In Column 4, it uses the shrunken leave-two-out mean among all teachers with missing VA predictions. In column 5, it uses the shrunken leave-two-out mean among all teachers at the school with missing VA predictions. All specifications include school-year fixed effects. N=79,466 school-grade-subject-year cells in Column 1; 91,221 in Columns 2-5 in Panel A; and 90,701 in Columns 2-5, Panels B-C.

Supplement Table 2. Robustness of CFR-I, Table 5's robustness results
Quasi-Experimental Estimates of Forecast Bias: Robustness Checks

	Teacher Exit Only		Full Sample		<25% Imputed VA		0% Imputed VA	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A: Quasi-experimental models without controls							
Change in mean teacher predicted VA	1.174	1.080	0.936	0.904	1.100	0.965	1.081	0.918
	(0.040)	(0.044)	(0.022)	(0.022)	(0.035)	(0.040)	(0.043)	(0.051)
Year fixed effects	X		X		X		X	
School-year fixed effects		X		X		X		X
Number of School x Grade x Subject x Year Cells	79,466	79,330	91,221	91,221	34,495	34,495	23,445	23,445
	Panel B: Models for change in prior-year scores							
Change in mean teacher predicted VA	0.296	0.226	0.175	0.093	0.199	0.064	0.177	0.033
	(0.039)	(0.043)	(0.023)	(0.022)	(0.033)	(0.038)	(0.040)	(0.047)
	Panel C: Models for change in end-of-year scores, with controls for change in prior-year scores							
Change in mean teacher predicted VA	0.981	0.928	0.853	0.859	0.978	0.926	0.973	0.899
	(0.030)	(0.029)	(0.019)	(0.017)	(0.028)	(0.031)	(0.035)	(0.041)
Change in mean student prior year score	0.650	0.675	0.497	0.537	0.611	0.608	0.610	0.583
	(0.004)	(0.005)	(0.009)	(0.009)	(0.006)	(0.007)	(0.007)	(0.009)

Notes: See notes to CFR (2014a), Table 5. Columns 1, 3, 5, and 7 in Panel A reproduce results from that table. Even-numbered columns add school-year fixed effects. Panel B changes the dependent variable, while Panel C adds a control for the change in the prior-year score.